Towards Multimodal Domain Generalization with Few Labels

Supplementary Material

1. Training Details

Data Augmentation. To make effective use of unlabeled data through consistency regularization, we design a modality-specific augmentation pipeline that clearly separates weak and strong perturbations for video, audio, and optical flow.

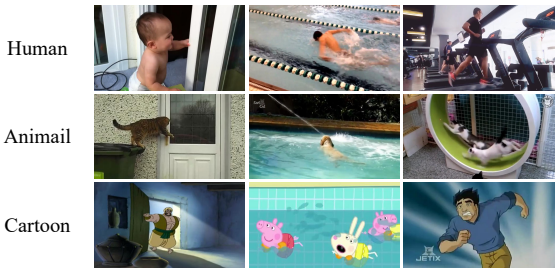
- Video: Weak augmentation includes random horizontal flips and spatial translations. Strong augmentation applies RandAugment [7] followed by Cutout [9] to introduce more substantial appearance variations.
- Audio: Weak augmentation consists of random gain changes and pitch shifts. Strong augmentation adopts SpecAugment [21] (frequency and time masking) together with additive noise to simulate acoustic variations and corruptions common in diverse domains.
- Optical Flow: We follow a strategy analogous to the video pipeline. Weak augmentation uses random flips and translations, while strong augmentation applies Cutout [9] and noise injection to the flow fields to prevent overfitting to specific motion artifacts.

Network Architectures. Our framework is implemented using the MMAction2 [6] toolkit. The specific encoders for each modality are as follows:

- Video Encoder: We employ a SlowFast network [14] pre-trained on Kinetics-400 [17]. The output feature dimension is $d_v = 2304$.
- Audio Encoder: We utilize a ResNet-18 [16] architecture, pre-trained on VGGSound [3]. The output feature dimension is $d_a = 512$.
- Optical Flow Encoder: We use a SlowFast network [14] configured with a slow-only pathway, initialized with Kinetics-400 [17] weights. The feature dimension is $d_f = 2048$.
- Cross-Modal Translators: To enable the CMPA module, the translators $(t_{v\to a}, t_{a\to v})$ are modeled as two-layer Multi-Layer Perceptrons (MLP) with 2048 hidden units and ReLU activation, projecting features between modality-specific subspaces.

Optimization and Hyperparameters. We optimize the entire framework using the AdamW optimizer [20] with a base learning rate of 1×10^{-4}, a weight decay of 1×10^{-3}, and a batch size of 32. The loss balancing hyperparameters are set to $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$. The confidence threshold for pseudo-labeling is set to $\tau = 0.95$. The robustness parameter for the Generalized Cross-Entropy loss in the DAR module is set to $q = 0.7$. All experiments are conducted on two NVIDIA RTX 4090 GPUs.

(a) EPIC-Kitchens Dataset



(b) HAC Dataset

Figure 1. Visualization of domain shifts in the experimental benchmarks. (a) Example frames from the EPIC-Kitchens dataset across three environments (D1, D2, D3), highlighting variations in viewpoint and illumination. (b) Samples from the HAC dataset (Human, Animal, Cartoon), illustrating stylistic differences across domains.

2. Dataset Details

EPIC-Kitchens. We utilize the domain adaptation subset of EPIC-Kitchens [8], a large-scale egocentric video dataset. Following standard MMDG protocols [10], we organize the data into three domains (D1, D2, D3), representing three distinct kitchen environments, as shown in Fig. 1 (a). These domains differ in terms of lighting, spatial layout, and background clutter. The dataset contains 8 overlapping action classes: 'put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour'. This setup evaluates the model's ability to generalize egocentric action recognition across environmental changes.

HAC. The Human-Animal-Cartoon (HAC) dataset [10] is employed to evaluate domain generalization under stylistic shifts. The dataset comprises three visually disjoint domains: Human (H), Animal (A), and Cartoon (C), as illustrated in Fig. 1 (b). This setup challenges the model to learn action semantics that are invariant across appearance discrepancies and kinematic variations. The dataset consists of 7 common action classes: 'sleep', 'watch TV', 'eat', 'drink', 'swim', 'run', and 'open door'.

3. Implementation Details of Baselines

To ensure a fair comparison, all baseline methods are re-implemented using the MMAction2 [6] codebase. We strictly maintain the same backbone architectures (SlowFast [14] for video, ResNet-18 [16] for audio). Consistent with our proposed framework, all methods are optimized using AdamW [20] with a base learning rate of 1×10^{-4} and a batch size of 32.

MMDG Baselines. These methods use only the labeled subset of source domains to learn domain-agnostic features.

- **RNA-NET [22]:** We incorporate the Relative Norm Alignment (RNA) loss as a regularization term alongside standard cross-entropy. By minimizing the discrepancy between feature norms of video and audio modalities relative to their class prototypes, RNA-NET aligns the modality distributions to improve generalization.
- **SimMMDG [10]:** Following the official implementation, we employ a decomposition head to disentangle features into modality-specific and modality-shared components. We apply supervised contrastive learning to the shared features to enforce semantic consistency, while applying distance constraints to the specific features to maintain modality distinctiveness.
- **MOOSA [11]:** We implement two self-supervised pre-tasks: Masked Cross-modal Translation and Multimodal Jigsaw Puzzles. These tasks are optimized jointly with the primary classification objective, encouraging the model to learn robust, domain-invariant representations.
- **CMRF [13]:** We implement Cross-Modal Representation Flattening by applying convex combination interpolation on multimodal feature representations. A distillation loss constrains the unimodal networks, enforcing a flatter loss landscape to mitigate the competitive imbalance between modalities.
- **MDJA [19]:** We implement Modality-Domain Joint Adversarial training by attaching domain discriminators to both unimodal and fused feature extractors. The model utilizes a Gradient Reversal Layer (GRL) to learn features that are discriminative for class boundaries but invariant to domain shifts.

SSL Baselines (Extended to Multimodal). We extend original unimodal SSL methods to the multimodal setting (denoted as ^{M}) by performing late fusion on features before applying the respective semi-supervised strategies.

- **FixMatch^M [23]:** We apply weak and strong augmentations to all input modalities. A pseudo-label is generated from the fused prediction of the weakly augmented view. This pseudo-label supervises the fused prediction of the strongly augmented view, provided the prediction confidence exceeds a fixed threshold ($\tau = 0.95$).
- **FreeMatch^M [25]:** We replace the fixed threshold of FixMatch with Self-Adaptive Thresholding (SAT). The threshold is dynamically adjusted based on the exponential moving average (EMA) of the model's confidence on unlabeled data, improving the utilization of hard classes and noisy domains.
- **CGMatch^M [5]:** CGMatch utilizes the Count-Gap (CG) metric to filter high-value unlabeled samples. We compute the CG score using multimodal predictions and apply the Filtering-by-Dynamic-Selection rule to retain reliable pseudo-labeled samples, strictly following the authors' open-sourced training recipes.

SSML Baselines. These methods are inherently designed to handle multimodal data dynamics under label scarcity.

- **Co-training [2]:** Treating video and audio as distinct views, we train two independent classifiers on the labeled data. In an iterative process, high-confidence predictions from the video classifier generate pseudo-labels for the Audio classifier, and vice-versa, leveraging the independence of the modalities.
- **TCGM [24]:** We implement Total Correlation Gain Maximization by adding an information-theoretic loss term to the objective. This maximizes the mutual information (Total Correlation) between video and audio modalities, utilizing unlabeled data to uncover latent semantic structures shared across views.
- **MSC [4]:** We implement the Strategic Complementarity learning mechanism. The model dynamically weights the contribution of each modality's prediction based on estimated reliability and consistency, allowing the stronger modality to rectify the weaker one during semi-supervised training.
- **DECPL [1]:** Adopting the Audio-Visual Contrastive learning paradigm, we apply a contrastive loss between audio and video representations alongside consistency regularization. This ensures semantic alignment across modalities is maintained even in the absence of ground-truth labels.
- **STiL [12]:** Originally designed for tabular-image data, we adapt STiL for the Video-Audio setting. We decompose multimodal representations into shared and specific subspaces via disentangled contrastive consistency and use a consensus-based pseudo-labeling strategy to filter unlabeled samples, refining labels via prototype-guided smoothing.

SSDG Baselines (Extended to Multimodal). We adapt methods designed for semi-supervised domain generalization to handle multimodal inputs, focusing on cross-domain robustness.

- **StyleMatch^M [26]:** We extend the stochastic style consistency strategy to the multimodal. We apply stochastic style transfer augmentation to video and audio inputs to simulate domain shifts. A consistency loss is then enforced between the predictions of the original and perturbed multimodal inputs.
- **UPUD^M [18]:** We implement the Unlabeled Proxy-based

Contrastive (UPC) and Surrogate Class (SC) components to better exploit low-confidence unlabeled samples. We extend the method to the multimodal scenario by applying the UPC and SC objectives on the fused video-audio embeddings.

- **NIED-LR^M [15]:** We integrate the Domain-Guided Weight Modulation (DGWM) module into a FixMatch backbone. Using available source domain labels, DGWM modulates classifier weights to be domain-aware, thereby enhancing cross-domain generalization capabilities.

References

[1] Maregu Assefa, Wei Jiang, Jinyu Zhan, Kumie Gedamu, Getinet Yilma, Melese Ayalew, and Deepak Adhikari. Audio-visual contrastive and consistency learning for semi-supervised action recognition. *IEEE Transactions on Multimedia*, 26:3491–3504, 2024. 2

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1

[4] Junchi Chen, Richong Zhang, and Junfan Chen. Semi-supervised multimodal classification through learning from modal and strategic complementarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15812–15820, 2025. 2

[5] Bo Cheng, Jueqing Lu, Yuan Tian, Haifeng Zhao, Yi Chang, and Lan Du. Cgmatch: A different perspective of semi-supervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15381–15391, 2025. 2

[6] MMAction Contributors. Openmmlab's next generation video understanding toolbox and benchmark. 2020. 1, 2

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[10] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36:78674–78695, 2023. 1, 2

[11] Hao Dong, Eleni Chatzi, and Olga Fink. Towards multimodal open-set domain generalization and adaptation through self-supervision. In *European Conference on Computer Vision*, pages 270–287. Springer, 2024. 2

[12] Siyi Du, Xinzhe Luo, Declan P O'Regan, and Chen Qin. Stil: Semi-supervised tabular-image learning for comprehensive task-relevant information exploration in multimodal classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15549–15559, 2025. 2

[13] Yunfeng Fan, Wenchao Xu, Haozhao Wang, and Song Guo. Cross-modal representation flattening for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 37:66773–66795, 2024. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 2

[15] Chamuditha Jayanaga Galappaththige, Zachary Izzo, Xilin He, Honglu Zhou, and Muhammad Haris Khan. Domain-guided weight modulation for semi-supervised domain generalization. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6495–6505. IEEE, 2025. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[18] Dongkwan Lee, Kyomin Hwang, and Nojun Kwak. Unlocking the potential of unlabeled data in semi-supervised domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30599–30608, 2025. 2

[19] Hongzhao Li, Hualei Wan, Liangzhi Zhang, Mingyuan Jiu, Shupan Li, Mingliang Xu, and Muhammad Haris Khan. Towards robust multimodal domain generalization via modality-domain joint adversarial training. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 180–188, 2025. 2

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2

[21] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 1

[22] Mirco Planamente, Chiara Plizzari, Simone Alberto Peirone, Barbara Caputo, and Andrea Bottino. Relative norm alignment for tackling domain shift in deep multi-modal classification. *International Journal of Computer Vision*, 132(7): 2618–2638, 2024. 2

[23] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2

[24] Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *European conference on computer vision*, pages 171–188. Springer, 2020. 2

[25] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *Eleventh International Conference on Learning Representations*. OpenReview. net, 2023. 2

[26] Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, 131 (9):2377–2387, 2023. 2